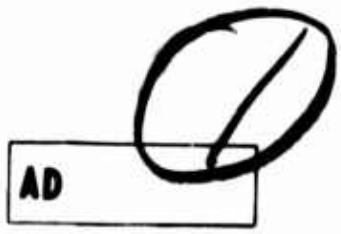
Technical Research Note 214



CHECKER CONFIDENCE STATEMENTS AS AFFECTED BY PERFORMANCE OF INITIAL IMAGE INTERPRETER

Michael G. Samet

SUPPORT SYSTEMS RESEARCH DIVISION

U. S. Army
Behavioral Science Research Laboratory

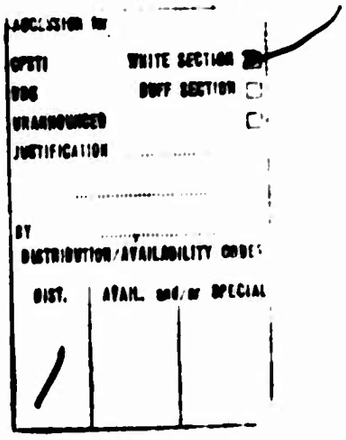
September 1969

35

BEHAVIORAL SCIENCE RESEARCH LABORATORY

An activity of the Chief, Research and Development

J. E. UHLANER
Director



NOTICES

DISTRIBUTION: Primary distribution of this report has been made by BESRL. Please address correspondence concerning distribution of reports to: U. S. Army Behavioral Science Research Laboratory, Attn: CRDBSRL, Room 239, Commonwealth Building, 1320 Wilson Blvd., Arlington, Virginia 22209.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Behavioral Science Research Laboratory.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Technical Research Note 214

AD

CHECKER CONFIDENCE STATEMENTS AS AFFECTED BY PERFORMANCE OF INITIAL IMAGE INTERPRETER

Michael G. Samet

Robert Sadacca, Task Leader

SUPPORT SYSTEMS RESEARCH DIVISION
Joseph Zeidner, Chief

U. S. ARMY BEHAVIORAL SCIENCE RESEARCH LABORATORY

Office, Chief of Research and Development
Department of the Army

Room 239, The Commonwealth Building
1320 Wilson Boulevard, Arlington, Virginia 22209

September 1969

Army Project Number
2Q662704A721

Image Systems a-00

BESRL Technical Research Reports and Technical Research Notes are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The SURVEILLANCE SYSTEMS research program of the U. S. Army Behavioral Science Research Laboratory has as its objective the production of scientific data bearing on the extraction of information from surveillance displays, and the efficient storage, retrieval, and transmission of this information within an advanced computerized image interpretation facility. Research results are used in future systems design and in the development of enhanced techniques for all phases of the interpretation process. Research is conducted under Army RDT&E Project No. 2Q662704A721, "Surveillance Systems: Ground Surveillance and Target Acquisition Interpreter Techniques," FY 1969 Work Program.

The development of procedures to maintain and improve the proficiency of image interpreters within an image interpretation facility is one of the major objectives of the Work Unit, "Information Processing in Advanced Image Interpretation Systems--IMAGE SYSTEMS." The present publication reports on one aspect of assigning interpreters to work as two-man teams in which one interpreter checks interpretations made independently by his teammate. The study concentrates on the checker's statements of his confidence in identifications as affected by varying levels of identification accuracy and confidence validity on the part of the initial interpreter.

J. E. UHLANER, Director
U. S. Army Behavioral Science
Research Laboratory

CHECKER CONFIDENCE STATEMENTS AS AFFECTED BY PERFORMANCE OF INITIAL IMAGE INTERPRETER

BRIEF

Requirement:

Research to develop effective team procedures for image interpretation requires study of the type and amount of information exchanged among team members. The objective of the present study was to determine how an initial interpreter's accuracy of identification and validity of stated confidence in his identifications affect the usefulness of the checker's confidence statements.

Procedure:

Identifications of 60 annotated targets and associated confidence statements were obtained from 18 newly trained image interpreters. Confidence estimates were stated under a point payoff scheme in which it was to the disadvantage of the interpreter to overstate or understate his confidence. Half of the interpreters were given individual performance feedback. Interpreters were then presented with three sets of 60 annotated images to which identifications and confidence statements attributed to an initial checker were attached. The information provided incorporated three levels of identification accuracy and three levels of confidence validity, arranged according to a Graeco-Latin square research design. The task of the interpreter was to examine each annotation, note the previous identification and confidence statement, and then state his own confidence in the identification. The payoff scheme used in the preliminary set was applied.

Findings:

1. Checkers typically improved on the confidence validity of interpreters who were poor or only moderately good in stating confidence.

2. Checker confidence statements in identifications made by interpreters with an "excellent" record of confidence validity were less valid than those of the initial interpreters.

3. Interpreters' confidence statements were more valid when they were checking than when they were stating confidence in their own identifications.

4. Checkers' confidence statements were more affected by observed variations in the identification accuracy of the initial interpreter than by his confidence validity.

5. Knowledge of their own initial identification and confidence proficiency did not affect checker performance.

Utilization of Findings:

In team operations, confidence statements made by initial interpreters who have excellent records of estimating the probability that their identifications are correct should be allowed to stand.

The checkers' confidence statements are to be preferred when initial statements are supplied by interpreters whose past performance in making such statements is poor or only moderately good.

CHECKER CONFIDENCE STATEMENTS AS AFFECTED BY PERFORMANCE OF INITIAL IMAGE INTERPRETER

CONTENTS

Page

TABLES

FIGURES

CHECKER CONFIDENCE STATEMENTS AS AFFECTED BY PERFORMANCE OF INITIAL IMAGE INTERPRETER

Image interpreter teams in which one man checks the reports of another have generally yielded more accurate and complete information than the average interpreter working alone.[1] However, in some instances, teams have not shown any improvement.[2] An individual interpreter's performance may even suffer as a result of his being part of a team. It was found that the better the initial interpreter, the less the improvement resultant from adding a checking interpreter; and conversely, the better the checking interpreter, the greater the improvement.[3]

Experiments on interpreter/checker performance have generally concentrated on completeness, accuracy, and timeliness as measures of individual and team proficiency. With the advent of computerized intelligence systems, emphasis on techniques for processing probabilistic information has grown. Recognition that interpreter identifications of dispersed and concealed enemy targets can seldom be made with 100% certitude has led to study of the interpreter's confidence in his findings and its operational use in assessing the probability that given identifications are in fact correct. Use of the initial interpreter's confidence in his identifications to help determine which identifications a checker should examine has been explored.[4] However, the direct effect of the initial interpreter's confidence on the checker's own accuracy and confidence has not been systematically studied. In view of the impact that suggestive information can have on interpreter performance, a study was undertaken to evaluate checker performance when the intelligence information he is checking is furnished by interpreters with varying records of accuracy of identifications and confidence statements.

---

[1] Sadacca, R., H. Martinek, and A. I. Schwartz. Image Interpretation Task--Status Report. Technical Research Report 1129. U. S. Army Behavioral Science Research Laboratory. June 1962.

[2] Bolin, S. F., R. Sadacca, and H. Martinek. Team procedures in image interpretation. Technical Research Note 164. U. S. Army Behavioral Science Research Laboratory. December 1965.

[3] Doten, G. W. and R. Sadacca. Team interpretation procedures: Selection of teammates and role assignment. Technical Research Note 201. U. S. Army Behavioral Science Research Laboratory. January 1969.

[4] Doten, G. W., J. T. Cockrell, and R. Sadacca. The use of teams in image interpretation: information exchange, confidence, and resolving disagreements. Technical Research Report 1151. U. S. Army Behavioral Science Research Laboratory. October 1966.

SPECIFIC OBJECTIVE

The specific objective of the present study was to determine how different levels of identification accuracy and of confidence validity associated with an initial interpreter affect the confidence validity of the checker. Of secondary interest was whether checkers supplied with some knowledge of their own prior identification and confidence performance would be affected differently.

METHOD

Experimental Subjects

Eighteen interpreters recently graduated from the image interpretation course at the U. S. Army Intelligence School, Fort Holabird, Maryland, participated as subjects. All had met the school's entrance requirement of a score of 100 or above in the General Technical Aptitude Area (composite of Verbal and Arithmetic Reasoning tests).

Experimental Materials

Stimulus imagery consisted of four sets of 60 annotated targets. Sets were carefully matched on target type, photo quality, scale, and level of concealment; for example, each set had exactly the same number of 3/4-ton trucks at good image quality, 1:5000 scale, and partial concealment. However, the ordering of target type within each imagery set depended upon the position of the targets in the roll of imagery and was not identical for all sets. A list of potential target names as given to the interpreters appears in Figure 1.

TARGET LIST

Armored Personnel Carrier (APC)

Howitzer SP-105
Howitzer SP-155
Howitzer Towed - 105
Howitzer Towed - 155

Shelter - Canvas

Tank - M-41
Tank - M-48

Pup Tent
Command Post (CP) Tent
General Purpose (GP) Tent - Medium
General Purpose (GP) Tent - Large

Cargo Trailer 1/4
Cargo Trailer 3/4
Cargo Trailer 1½
Water Trailer 1½
Ammunition Trailer 2
Semi-Trailer (low Bed)

Truck 1/4
Truck 3/4
Truck 2½
Dump Truck 2½
Truck 5

Bulldozer
Tractor
Wrecker
Civilian Vehicle

Radar Antenna

Figure 1. Target List

Experimental Design

The four sets of imagery, assumed to be equivalent, were randomly assigned to each task requirement. One set was used in an initial performance test which required subjects to supply identifications and confidence statements for each annotation. Target identifications were developed for each of the three remaining sets of imagery. Of the 60 annotations in each set, percentages correct were 25%, 50%, and 75%. Misidentifications usually named a target type likely to be confused with the target shown. Next, a hypothetical statement of confidence was assigned to each identification. Each set of confidence estimates included six at 50% and three at all other 5% steps ranging from 5% to 95% (60 confidence values in all).

The three sets of confidence statements were associated with appropriate target identifications to provide three levels of confidence validity--square of the biserial coefficient of correlation between confidence and accuracy of identification--equal to .00, .40, and .80. Nine sets of identifications and associated confidence statements were thus generated to represent all combinations of the three levels of identification accuracy and the three levels of confidence validity. Experimental conditions for each subject were fixed by random assignment to a row position in a 3 x 3 Graeco-Latin square (Figure 2).

Procedure

Preliminary Phase. To obtain individual measures of base performance for comparison with team performance, interpreters in the sample were asked to identify the annotated targets in the first set of 60 targets and to state their confidence in each identification. They were instructed to state their feeling of confidence in light of a special payoff scheme designed to discourage overstatement or understatement of their actual confidence. The values used are shown in Table 1; a 100-point penalty was threatened for each misidentification. The payoff function and rationale for use of the procedure in the present study are explained in Appendix B. The integration of the payoff scheme into the experimental procedure is elaborated in the instructions to interpreters (Appendix A).

After the preliminary phase, half the subjects were given a key to the ground truth of the annotated targets and asked to review each annotation, marking each of their identifications as correct or incorrect. They were then individually shown how effectively they had used the payoff scheme in accordance with their own responses. The other half received no feedback.

Row	Number of Subjects	Feedback After Preliminary Phase	SESSION Experimental Phase		
			I Identification Accuracy / Confidence Validity	II Identification Accuracy / Confidence Validity	III Identification Accuracy / Confidence Validity
1	3	Yes	25%/.00	50%/.40	75%/.80
	3	No			
2	3	Yes	50%/.80	75%/.00	25%/.40
	3	No			
3	3	Yes	75%/.40	25%/.80	50%/.00
	3	No			

Figure 2. Design of Experiment

Experimental Phase. In the experiment proper, each interpreter was given three sets of annotated imagery. With each set he was given identifications and associated confidence statements attributed to a previous interpreter. He was instructed to examine each annotated target and check the initial identification and associated confidence statement and then to state his own confidence in the given identification. He was not to modify the identification. He was told that the point payoff scheme used in the preliminary phase of the experiment would also be invoked here to evaluate his own confidence statements and to compare them with those of the initial interpreter. Instructions to the interpreters appear in Appendix A.

Dependent Variables

Listed below are seven measures derived from the data and used in the analyses. Each was computed separately for each set of 60 responses. The first five measures are also meaningfully defined across the responses of each of the nine hypothetical initial interpreters, with initial interpreter values for identification accuracy and confidence validity serving as the principal independent variables. Variables 6 and 7 measure the relationship between performance of the initial interpreter and that of the checker.

1. Identification accuracy. Number of correct (to target type and model or size) identifications divided by the total number of identifications (60) expressed as per cent.

2. Confidence validity. Square of biserial correlation between confidence statement and correctness of the identification.

3. Point score. Mean number of payoff points achieved by interpreter.[5]

4. Inappropriate confidence. Number of times confidence in an incorrect identification was greater than 50% plus number of times confidence in a correct identification was less than 50%.

5. Confidence spread. Tendency to make very high or very low confidence statements.

$$\text{Formula:} \quad \frac{\sum_{i=1}^{60} (c_i - 50)^2}{60}$$

where c_i = confidence that identification i is correct.

[5] No points were actually subtracted for incorrect identification as threatened in the preliminary phase.

Table 1

PAYOFF VALUES USED IN EXPERIMENT

Level of Confidence that Identification is Correct	Point Credits if Identification is Correct	Point Credits if Identification is Incorrect
%	C	I
100	100	0
95	99	10
90	98	19
85	97	28
80	96	36
75	94	44
70	91	51
65	88	58
60	84	64
55	80	70
50	75	75
45	70	80
40	64	84
35	58	88
30	51	91
25	44	94
20	36	96
15	28	97
10	19	98
5	10	99
0	0	100

6. Checker/initial interpreter relationship. Correlation between checker's confidence statements and those attributed to initial interpreters (z-transformation).

7. Checker/initial interpreter relationship with accuracy of identification partialed out (z-transformation).

RESULTS AND INTERPRETATION

Table 2 summarizes preliminary phase data for interpreter performance without knowledge of "previous" identifications or confidence statements. Data were analyzed to describe differences in performance between average experimental interpreter and hypothetical initial interpreters and to get an idea of relationships among dependent variables. Mean identification accuracy of 37% falls between the first two levels of hypothetical interpreter identification accuracy, 25% and 50%, respectively. Mean confidence validity of .13 falls between the first two levels of hypothetical interpreter confidence validity, .00 and .40, respectively. These contrasts permit the established levels of hypothetical initial interpreter identification accuracy and confidence validity to be reasonably labeled (for future reference) as relatively poor, good, and excellent, respectively. Rather poor confidence performance during the preliminary phase is also reflected in the low mean point score (68) and high mean inappropriate confidence (24). In fact, had interpreters stated 50% confidence for every response they would have obtained a higher point score (75). A mean confidence spread of 993 indicates greater use of confidence values near 0 or 100 than was attributed to the hypothetical interpreters. In regard to the intercorrelations among variables, it is not surprising that confidence validity, point score, and inappropriate confidence intercorrelated significantly since all three were intended to measure the degree of correspondence between confidence and ground truth. From the significant correlation coefficients obtained for each of these three variables with identification accuracy, subjects with superior identification accuracy also gave superior confidence performance.

An analysis of variance was performed on each dependent variable in the experimental phase, and the significant F-ratios are given in Table 3. For no variable did feedback at the end of the preliminary phase prove to be a significant effect. For all dependent variables but one, significant differences were found only for the main effects of major interest: identification accuracy/imagery set and confidence validity. (Because of the care taken to match imagery sets, the identification accuracy/imagery set effects are assumed to be due mainly to differences in identification accuracy as opposed to imagery variations.) Mean initial levels of identification accuracy and confidence validity for the checkers are presented in Table 4 in comparison with the means established for the three initial interpreters at each level of identification accuracy and confidence validity.

Table 2

MEANS AND INTERCORRELATION COEFFICIENTS
FOR PRELIMINARY PHASE
(N = 18)

Variable	M	Intercorrelation			
		Confidence Validity	Point Score	Inappropriate Confidence	Confidence Spread
Identification Accuracy	35.85	.525*	.562*	-.587*	.408
Confidence Validity	.17	-	.728**	-.745**	.163
Point Score	68.03	-	-	-.877**	-.256
Inappropriate Confidence	24.05	-	-	-	-.061
Confidence Spread	993.26	-	-	-	-

*Significant Correlation, P < .05.
**Significant Correlation, P < .01.

On the three measures of confidence performance--confidence validity, point score, and inappropriate confidence--checkers were generally able to improve more substantially on mean initial interpreter performance in the case of a poor or good interpreter than in the case of an excellent interpreter. Checker improvement over initial interpreter performance showed a general decline as the level of initial performance increased. In fact, a degradation in performance was observed for confidence validity at the highest level of initial confidence validity and for point score at the highest levels of both initial identification accuracy and confidence validity. However, checker means for confidence validity, point score, and inappropriate confidence indicated far superior performance than was observed for the same variables during the preliminary phase. Checker performance on these variables was better at the higher levels of initial confidence validity; however, performance was best at the lowest level of initial identification accuracy, 25%. Although the analysis of variance design did not permit the recovery of a term for interaction between initial identification accuracy and confidence validity, intuition would suggest that some kind of interaction was present.

By design, confidence spread was identical across the hypothetical initial interpreters and equal to 712.5. Checker confidence spread decreased with increasing initial identification accuracy, indicating that checkers made more extreme confidence statements when reviewing the responses of a less accurate interpreter. Mean confidence spread during checking was far larger than when interpreters were assigning confidence to their own identifications. Interpreters were apparently more willing to state extreme confidence in an identification made by someone else.

The obtained relationships between confidence performance of checkers and hypothetical initial interpreters are shown in Table 5. Mean values suggest greater acceptance of initial confidence statements when the rate of identification accuracy was observed to be more or less distinct (25% or 75%) than when observed to be chance (50%). Although considerably lower, partial relationship between checker and initial interpreter remained significantly different and in the same direction across the three initial accuracy levels. Checker confidence validity clearly increased with initial interpreter confidence validity; however, that the increase was for the most part attributable to the perceived correctness or incorrectness of the identifications is shown by the nonsignificant differences obtained when accuracy of identification was partialed out. Checkers in stating their own confidence were generally more influenced by the initial interpreter's overall accuracy rate than by the impact of his confidence for an individual target identification.

Of supplementary interest is the finding that checker/initial interpreter partial relationship values decreased significantly after the first checking session: session means were .335, .178, and .188 ($P < .05$). Thus, checker tendency to rely on initial interpreter confidence declined with task experience.

Overall, the results point to the following general explanation: Checkers tended to augment the initial interpreter's confidence value when the checker perceived the identification to be correct and to reduce it when he perceived it to be incorrect; in each case, however, the checker exercised temperance. The amount of temperance was far more for perceived correct identifications than for perceived incorrect identifications. That is, checkers tended to use a more extreme confidence statement when in disagreement with the identification.

Table 3

SIGNIFICANT F-RATIOS FOR DEPENDENT VARIABLES

Source of Variation	DF	Confidence Validity	Point Score	Inappropriate Confidence	Confidence Spread	Checker/Initial Interpreter	
						Relationship	Partial Relationship
Between Subjects							
Rows (R)	2						
Feedback (F)	1						
R x F	2						
Subj w/Gp. (Error Between)	12						
Within Subjects							
Sessions (S)	2						5.0*
Identification Accuracy/Imagery (A)	2	5.8**	58.8**	24.8**	25.0**	3.8*	5.8**
Confidence Validity (B)	2	6.9**	5.4*			29.9**	
SF	2						
AF	2						
BF	2						
S x Subj w/Gp. (Error Within)	24						

*P < .05
**P < .01

Table 4

MEANS FOR HYPOTHETICAL INITIAL INTERPRETERS AND CHECKERS

Dependent Variable		Independent Variable: Initial Identification Accuracy				Independent Variable: Initial Confidence Validity			
		25%	50%	75%	Significance Between Checker Column Means	.00	.40	.80	Significance Between Checker Column Means
Confidence Validity	Initial Interpreter	.412	.353	.428		.000	.400	.800	
	Checker	.710	.506	.487	P < .05	.419	.607	.676	P < .01
Point Score	Initial Interpreter	76.5	78.2	76.8		67.9	79.7	83.9	
	Checker	87.2	79.4	74.8	P < .01	78.3	81.2	81.8	P < .05
Inappropriate Confidence	Initial Interpreter	18.7	16.0	19.3		27.7	14.3	12.0	
	Checker	8.0	12.8	16.2	P < .01	13.7	11.6	11.7	n.s.
Confidence Spread	Initial Interpreter	712.5	712.5	712.5		712.5	712.5	712.5	
	Checker	1696.8	1564.6	1357.4	P < .01	1505.4	1584.7	1528.8	n.s.

Table 5

MEANS FOR VARIABLES MEASURING RELATIONSHIP BETWEEN CHECKER AND INITIAL INTERPRETERS' CONFIDENCE

Checker/Initial Interpreter Relationship	Independent Variable							
	Initial Identification Accuracy			Significance between Column Means	Initial Confidence Validity			Significance between Column Means
	25%	50%	75%		.00	.40	.80	
Z-transform of Correlation Coefficient	.451	.342	.492	P < .05	.191	.480	.613	P < .01
Z-transform of Partial Correlation Coefficient	.239	.136	.325	P < .01	.227	.239	.234	n.s.

CONCLUSIONS AND DISCUSSION

The conclusions of principal interest for team interpretation methods are:

1. A checker can usually improve on the confidence validity of an initial interpreter who is relatively poor or good in making confidence statements, but most checkers will degrade the confidence validity of an excellent confidence assessor.

2. The confidence validity of an interpreter when he is performing a checking function is considerably above the validity of his confidence in his own identifications.

3. In general, in checking confidence statements, a checker is more sensitive to initial interpreter variations in identification accuracy than to variations in confidence validity.

The first finding is consonant with results from other team method studies which indicate that the better the initial interpreter the less the gain can be expected through employing a team method. However, the low mean initial confidence validity of .17 obtained in the preliminary phase of the experiment indicates that the confidence statements assigned by most interpreters could stand considerable improvement. That checkers were less sensitive to confidence validity than to identification accuracy rates is not surprising considering the greater emphasis placed on accuracy on the job and in training. A "halo" effect may also be in operation. Perceiving the initial interpreter's accuracy rate to be no better than chance, the checker may tend to ignore the validity of his confidence statements.

Of secondary interest is the finding that the extra training afforded the interpreters receiving feedback apparently did not have any effect on subsequent performance. The payoff scheme represented a new response mode for the interpreters. Because of the short task duration, it is doubtful that more than a few came to understand its operation. Giving feedback after each response and not after a large block of responses as was done here might have had greater impact.

APPENDIXES

APPENDIX A

INSTRUCTIONS TO INTERPRETERS FOR PRELIMINARY PHASE OF EXPERIMENT

In the first phase of this experiment, you are to examine 60 annotated images all of which are actual targets. Your task is to identify each target. On a separate sheet is printed a list of targets from which you can choose; the target name you assign must appear on this list. However, please note that this is a general list of "possible" targets. Some of the items listed may not be among those you will be looking at. It is very important that you include in your identification the type and/or size of the target when more than one type or size appears on the target list. To illustrate, the response "Tank" will not be accepted; it must be "Tank - M-41" or "Tank - M-48." "Truck" is not acceptable; it must be, for example, "Truck 1/4," "Dump Truck 2 1/2," etc.

In addition to the identification we would like to know how confident you are that your identification is correct. You are to use a confidence scale that runs from 0 to 100, where 100 indicates that you are certain your identification is correct. If you use this scale accurately, all of the identifications for which you indicate 100% confidence should be correct, 80% of the identifications for which you indicate 80% confidence should be correct, 50% of the identifications for which you indicate 50% confidence should be correct, and so forth. You can use 0, 05, 10, 15, 20, ... 75, 80, 85, 90, 95, 100 to indicate your estimate of the probability that you have made a correct identification.

From previous experiments we have found that an interpreter's statements of confidence in his identifications are very important in evaluating the accuracy of his identifications; so try to be as accurate as possible. To help prevent you from over- or underestimating your degree of confidence, we are going to use a table of payoff credits specially designed to score the appropriateness of your confidence measures. If you look at the payoff sheet, you see three separate columns. In the first column of the table are listed confidence levels from 100 to 0 at 5% intervals. In the second and third columns are listed--corresponding to the confidence level--the number of credits or points you will win if the particular identification being judged is correct, and the number of points you will win if it is incorrect. You may observe that the more confident you are that a given identification is correct, the more points you will win if it indeed is CORRECT, and the less confident you are that it is correct the more points you will win if it indeed is INCORRECT. For example, if you are 100% confident of an identification, you will get 100 points if it is correct but you will get nothing if it is wrong. If you are 75% confident, you will get 94 points if you are correct and 44 points if you are incorrect. When you are 50% sure about an identification, you imply that you are equally confident of being correct as you are of being wrong. Therefore, at the 50% level of confidence you will get the same number of points whether you are right or wrong, namely 75. If you are 25% confident, you will get 44 points if you are correct and 94 points if you are incorrect.

Notice that when you are 0% confident about an identification and you prove wrong the payoff table says you are entitled to the maximum number of points--100. But we are interested in the accuracy of your identification as well as in your ability to estimate confidence. Therefore, in Phase I of this experiment, for every <u>incorrect</u> identification you will be penalized 100 points. If you are 0% confident about a wrong identification, you will get 100 points according to the payoff but will lose 100 for being wrong so you will wind up with no points at all. It should therefore be clear that you have absolutely <u>nothing</u> to gain if you misidentify targets and assign a low probability to the misidentification.

In summary, the more honest you are about your level of confidence, the more points you stand to win. The points that you accumulate for each identification will be summed and at the end of the experiment you will be provided with your total score. You will also be given a statement as to how well you did in comparison with the other interpreters who participated in the experiment. So please try to get as high a score as possible.

Blank responses are unacceptable. You must write down an identification for every annotation.

Are there any questions?

INSTRUCTIONS FOR PERFORMANCE FEEDBACK

Before giving you a key for the correct identifications, we want to give you some feeling for how well you are estimating the probability that your identifications are correct. The best way for us to accomplish this is to have you score your identifications and related confidence measures in accordance with the Table of Payoff Credits.

It is essential for the purposes of this research that you cooperate fully and honestly in scoring your own answer sheet. Also, you are expected to gain an accurate understanding of how the Payoff Table influences your level of confidence so as to allow you to become a better probability estimator in the sessions to follow. We will proceed as follows:

First, I will read off the list of correct identifications for the 60 annotations. Listen to each correct target name carefully, and then if your answer is correct mark a "C" in the column headed C/I; if your answer is incorrect mark an "I" in that same column. Your identification must be precisely correct. As an example, if the right answer is "truck - 3/4", you must have "truck - 3/4" to get a "C"; if you have listed "truck - 1/4", you get an "I". Let's do that now.

For each annotation, you now should have either a "C" or an "I". To score each response, look at the value for your confidence, find this value in column one of your Payoff Table, then select the corresponding number of point credits in column "C" if your response was correct or in column "I" if your response was incorrect. Write the resulting number in the column on your response sheet labeled PT. Do this for each one of the 60 responses. Return to your response sheets and make a small x to the right of every PT box for which the point value is less than 75. For every response which you now have an x to the right of the PT box, it means that you were either less than 50% confident of what turned out to be a correct response or more than 50% confident of what turned out to be an incorrect response. If you look at the Payoff Table for each x'ed response, you see how many points you won and how many you could have won if the outcome of your response had been more in line with your expressed level of confidence. Let us give some examples:

If you were 30% confident on what turned out to be a <u>correct</u> response, you only got 51 points, whereas you could have got 91 points if you had been 70% confident about the response; the difference is 40 points, which in this example represents the penalty you payed for <u>underestimating</u> your confidence. If you were 80% confident of what turned out to be an <u>incorrect</u> response, you got only 36 points when you could have got 96 points if you had been 20% confident about the response; the difference is 60 points, which in this example represents the penalty you payed for overestimating your confidence. Familiarize yourself with the impact of such point differences for every x'ed response, that is, for every response for which you greatly misjudged your level of confidence.

By carefully following our instructions for learning about the properties of the Payoff Table, it should become very clear to you that the best thing for you to do is to always respond with a confidence that honestly reflects how you feel about the particular annotation.

You are now to proceed, with the help of a key for the correct identifications, to re-examine each of the annotations. Pay special attention to those for which your response was marked with an x, that is, those for which your level of confidence was inappropriate.

You need not tally up your total point credits. We will do that for you, and at the same time we will subtract 100 points for each misidentification. You will get the results at the end of the experiment.

INSTRUCTIONS FOR EXPERIMENT PROPER

Your task now will be to examine a different set of 60 annotated targets from the same role of imagery. This set has already been interpreted by an image interpreter from a previous graduating class. For each annotation, this interpreter selected a few specific target names from the target list. The interpreter has also assigned to each target name a level of confidence that the target name is correct. For each annotation, you will be given one of the target names selected by the image interpreter together with the level of confidence assigned to it. It is very important that you understand that the target name listed is not necessarily the one which the interpreter thought was most probable. For example, if the interpreter was 30% confident that it was a 1/4 ton truck, he may have been 60% confident that it was a 1/4 ton trailer. Very often, however, the target name will have been his first choice. In fact, whenever the expressed level of confidence is greater than 50%, this means that the man was more confident of the listed target name than of any other. The 60 particular annotations that you will observe have been selected from a much larger set interpreted by the same man in a way that gives a good sample of the interpreter's confidence estimates.

Your task is as follows. Look at the annotated object and then at the identification and assigned confidence made by the previous interpreter. Then, in the appropriate space on the response sheet state your own confidence that the annotated target is in truth what the man reported it to be. In other words, if the man said it had probability of 50% of being an APC, tell us what you think the probability is that it is an APC. Your personal level of confidence may be similar to or very different from that of the previous man; it may be higher or lower. To use the same example, if you are very confident that the target in question is not an APC, then simply assign a very low probability to it. <u>You are always to estimate the probability that the specific annotation is actually the target identification listed on your response sheet</u>. You are not required to provide any alternative target names for any of the annotations. Since in this phase of the experiment you cannot make a misidentification, you will not be penalized as you were in Phase I. That is, 100 points will not be subtracted for any misidentification. However, your estimated confidences will be strictly scored according to the same payoff table employed in Phase I. Therefore, please try to be very accurate with your own confidence judgments.

After you complete the first set of 60 annotations in Phase II, you will be presented with another batch of 60 annotations together with a set of corresponding responses collected from those of a different image interpreter. The task procedure will be the same as explained above. Finally, you will be asked to respond in the same way to another set of similar image materials arranged from the responses of a third interpreter.

The man whose responses you are considering is identified by a number on the top of each respective response sheet. Be sure that you are working with a different man number in each of the three sessions here in Phase II. Remember that in all sessions your own confidence will be scored according to the payoff table. Your score will be compared to that achieved by the interpreter you are checking to see who was more accurate so try to estimate your confidence as accurately as possible.

After you complete a set of 60 annotations, please roll the imagery back to photo no. 1. For each man, you must examine the 60 annotations in order from 1 to 60; you are not permitted to go BACKWARDS.

Are there any questions?

NAME _____ DATE _____

#	Location Photo	Location Annotation	Identification	II Confidence	Your Confidence
1	02	06	Howitzer, SP-105	65	
2	02	16	Tlr, 3/4	85	
3	02	19	Trk, 2 1/2	60	
4	04	31	Trk, 2 1/2	95	
5	09	52	Trk, 1/4	95	
6	09	56	Pup Tent	15	
7	11	74	Trk, 3/4	80	
8	14	17	Tank, M-41	25	
9	16	26	Tlr, 1/4	50	
10	16	30	Tank, M-48	75	
11	17	35	Tank, M-48	65	
12	18	42	Trk, 1/4	10	
13	21	55	Trk, 1/4	50	
14	23	70	Tlr, 1/4	05	
15	24	02	Tank, M-48	60	
16	27	17	Trk, 2 1/2	70	
17	27	19	Tlr, 1 1/2	80	
18	27	20	Trk, 3/4	60	
19	27	21	Tlr,Tank,Wtr, 1 1/2	45	
20	28	30	Trk, 3/4	05	
21	32	34	APC	20	
22	34	47	Howitzer, SP-105	75	
23	36	69	Trk, 2 1/2	55	
24	38	04	Trk, 1/4	40	
25	39	24	Tank, M-48	25	

Figure A-1. Sample Response Sheet

APPENDIX B

RATIONALE FOR PAYOFF FUNCTION

After an image interpreter has identified a target, he is often asked to state his confidence that the identification is indeed correct. If his confidence statement is to have operational value, it is important that the statement accurately reflect his true feeling of confidence. In most experiments, the interpreter is simply asked to state a level of confidence, but this method can be criticized on grounds that there is no way of knowing if the stated confidence matches true confidence. For various implicit and/or explicit reasons that depend upon the personality of the interpreter and the given task, interpreters often tend to "hedge" their confidences; i.e., they may either overstate or understate true confidence if they see a particular advantage in doing so.

Several payoff schemes have therefore been developed to encourage honest statement of confidence (subjective probability)[1], [2]. If t is true confidence and c is stated confidence, then these functions are alike in that they grade a reward/penalty (usually points) for each response in accordance with a special nonlinear function of deviation of c from t. The quadratic payoff function[3] was instrumented in this study. The linear constraints ior the function were adjusted for convenience to make payoff credits positive with range from 0 to 100 (Table 1 of the text). For a correct identification, the interpreter was awarded $100-100(1 - c)^2$ points, but for an incorrect identification, he was rewarded $100-100c^2$ points.

It is necessary to show that in terms of normative decision theory, it is the subject's best strategy to always state his confidence accurately, i.e., to set c equal to t. As far as the interpreter is concerned, his expected number of points for any response is:

$$t \times [100 - 100(1 - c)^2] + (1 - t) \times [100 - 100c^2],$$

or

$$100 - 100\,(c^2 - 2tc + t).$$

---

[1] Toda, M. Measurement of subjective probability distribution. Institute for Research, Division of Mathematical Psychology. Report No. 3, 1963. State College, Pennsylvania.

[2] Roby, T. B. Belief states, evidence, and action. In Predecisional processes in decision making. USA Medical Research Laboratory Technical Document Report, No. 64-77, 1964, Behavioral Sciences Laboratory, Wright Patterson AFB.

[3] van Naerssen, R. F. A scale for the measurement of subjective probability. Acta Psychologica, 1962, 20, 159-166.

By taking the partial derivative with respect to c and setting it equal to zero, it follows that expected points will be a maximum if and only if c equals t. In summary, the closer c is to t, the more the interpreter has to gain expected point-wise.

During the preliminary phase of the experiment where interpreters made their own identifications, it would seem that the interpreter could take advantage of the payoff system by making intentional misidentifications and assigning very low confidence to them. For example, suppose the interpreter is fairly confident that the imaged object is some kind of vehicle. If he identifies it as a pup tent, knowing that this is clearly incorrect, but states 0% confidence, then he would get 100 points for an incorrect identification. To insure against this undesired possibility, the interpreter was told that he would be penalized 100 points for each incorrect identification (see instructions, Appendix A). However, as indicated in the definitions of dependent variables, no points were actually deducted in computing point score.

In addition to encouraging the subject to be honest, the payoff function served as a means of measuring confidence performance through mean number of points. Point score was obtained by employing the same payoff structure to score each confidence statement in light of whether the identification was correct or incorrect. Properties and uses of the quadratic and similar payoff functions as scoring rules have been discussed.[4]

The theory of admissible payoff functions for subjective probability measurement calls for a workable integration of mathematical[5] and psychological[6] constructs. The success of these measurement methods and the need for their incorporation into relevant experiments await further research. However, any fair test of the efficacy of the method would strive to adhere to the following task criteria:

1. The response mode and scoring method and their implications must be known and well understood by the interpreter. Training may be required to impress upon the interpreter the necessary correspondence between his own beliefs and the numbers into which these must be translated.[7]

---

[4] Winkler, R. L. The quantification of judgment: some methodological suggestions. Journal of American Statistical Association, 1967, 62, 1105-1120.

[5] Shuford, E. H., A. Albert and H. E. Massengill. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-147.

[6] de Finetti, B. Methods for discriminating levels of partial knowledge concerning 9 test items. British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

[7] Such training, although at a very superficial level, was attempted in the feedback mode after the preliminary phase of the present experiment.

2. The task must be so structured that it is to the disadvantage of the interpreter to respond in a manner inconsistent with his expectations. Maximization of expected points on each trial is to be achieved by making c congruent to t.

3. The interpreter should be keenly interested in maximizing his expected total score, each point added to the score having equivalent utility --either moral or material.

4. The method of measurement must be operational, efficient, and practical.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) U. S. Army Behavioral Science Research Laboratory, Arlington, Virginia	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

CHECKER CONFIDENCE STATEMENTS AS AFFECTED BY PERFORMANCE OF INITIAL IMAGE INTERPRETER

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

Michael G. Samet

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1969	37	4

8a. CONTRACT OR GRANT NO. b. PROJECT NO. DA R&D PJ No. 2Q662704A721 c. Image Systems Work Unit d. a-00	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Research Note 214 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office, Chief of Research and Development, DA, Washington, D. C.

13. ABSTRACT

One of the major objectives of the IMAGE SYSTEMS Work Unit is the development of procedures to maintain and enhance the proficiency of image interpreters within an advanced computerized image interpretation facility. Experiments on interpreter/checker performance have generally concentrated on completeness, accuracy, and timeliness as measures of individual and team proficiency. With the advent of computer-aided intelligence systems, emphasis on techniques for processing probabilistic information has grown. The present publication reports on one aspect of assigning interpreters to work as two-man teams in which one interpreter checks interpretations made independently by his teammate. The study was specifically concerned with determining how different levels of identification accuracy and of confidence validity associated with an initial interpreter affect the confidence validity of the checker. Four equivalent imagery sets of 60 annotated targets were used in the experimental procedure. The first set of 60 targets was assigned to 18 newly trained interpreters in an initial performance test which required the subjects to supply identifications and confidence statements for each annotation. Target identifications and confidence statements attributed to an initial checker were developed for each of the three remaining sets of imagery. In the preliminary test phase, confidence estimates were stated under a point payoff scheme in which it was to the disadvantage of the interpreter to overstate or understate his confidence. Half of the interpreters were given feedback on individual performance. In the experiment proper, each interpreter was given three sets of pre-annotated imagery with associated hypothetical confidence statements which he was required to examine, note previous identification/confidence information, and then state his own confidence in the given identification. Task performance was accomplished under the point payoff condition.

DD FORM 1 NOV 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
*Surveillance systems						
*Image Interpretation systems						
Information processing						
*Image Interpreter performance						
*Initial interpreters						
*Checkers						
*Confidence validity						
*Feedback techniques						
Laboratory facilities						
Target identification (classification)						
Image interpreter team performance						
Military psychology						
*Identification accuracy						
*Confidence estimates						
Checker/Interpreter relationship						